Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)
IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

An Intelligent System for Thyroid Disease Classification and Diagnosis

Aswathi A K
Dept. of Computer Science and Engineering
Jyothi Engineering College
Thrissur, India
aswathichittappurath@gmail.com

Anil Antony
Dept. of Computer Science and Engineering
Jyothi Engineering College
Thrissur, India
anil@jecc.ac.in

Abstract— **Data mining Techniques play a vital role in healthcare organizations such as for decision making, diagnosing disease and giving better treatment to the patients. Thyroid gland plays a major role in maintaining the metabolism of human body. Data mining in health care industry provides a systematic use of the medical data. Thyroid diseases are most common today. Early changes in the thyroid gland will not affect the proper working of the gland. By the early identification of thyroid disorders, better treatment can be provided in the early stage thus can avoid thyroid replacement therapy and thyroid removal up to an extent. This paper proposes a method for the classification and diagnosis of thyroid disease that a user is suffering from along with disease description and healthy advices. Support Vector Machine is used for classification. To optimize SVM parameters Particle Swarm Optimization is applied. User is provided with a window to enter the details such as the values of TSH, T3, T4 etc. There may be some values missing while the user entering the values. K-Nearest Neighbor algorithm is used for approximating the missing values in the user input**.

Keywords—data mining; thyroid disease; support vector machine; particle swarm optimization

I. INTRODUCTION

Data mining is the process of extracting large amount of data to identify and analyze data patterns. Data mining techniques can be used for discovering knowledge from large data base or knowledge base. So these techniques can help for extracting medical data patterns. This may help for analyzing the survivability of diseases. Medical data mining helps healthcare management, treatment effectiveness and patient involvement and relationships. Today thyroid disorders are common and are widespread worldwide. Thyroid has both structural and functional aspects. Thyroid is a butterfly-shaped gland which is located in the front of the neck. The hormone produced by thyroid gland plays a vital role in controlling human body system. The primary hormone produced by thyroid is Thyroxine (T4). A small portion of T4 is released from gland after releasing thyroxine to the blood stream and it is called triiodothyronine. Production of these hormones is controlled by TSH (Thyroid Stimulating Hormone), the hormone produced by pituitary gland. When T3 and T4 are more in the bloodstream, then pituitary gland releases less TSH and when they are less in the bloodstream, pituitary releases more TSH. Both the increase and decrease in thyroid hormone production may lead to health problems. Thyroid disease can either affects the function of the thyroid gland or can be a tumor. These abnormalities are caused by two reasons: production of too little thyroid hormone or production of too much thyroid hormone. Hyperthyroidism causes sudden weight loss, irregular heartbeats etc. Hypothyroidism causes thinning hair, heart attack etc. Goitre is another thyroid disorder that enlarges the thyroid gland. Early changes in the structural and functional aspects of the thyroid gland will not affect the proper working of the thyroid gland. So there is a frequent misunderstanding or misdiagnosing of the thyroid disorders. Thyroid disease classification by interpreting the values of the hormones is an important classification problem. Thyroid diseases are usually diagnosed by taking the values of TSH, T3, and T4 from the blood. This paper proposes a method for the classification and diagnosis of thyroid disorders using Weighted Support Vector Machine along with Particle Swarm Optimization to optimize SVM parameters.

II. LITERATURE SURVEY

K. Geetha and Capt. S. Santhosh Baboo in [3] have proposed a method to classify two major type of thyroid disease: Hyperthyroidism and Hyperthyroidism. In the pre-processing stage, missing values and not a number constraint are checked and missing values are filled by taking the mean value of the corresponding column. Then child subsets are created from the parent records using differential evolution algorithm. Subset of data is then applied to Kernel Based Bayesian Classification algorithm. Feature selection is performed using wrapper model. Here, the disease is classified into only two classes.

Jameel Ahmed and M. Abdul Rehman Soomrani in [1] have proposed a framework for diagnosing the thyroid disease type. The first phase is data pre-processing in which missing values in the dataset are filled using Medical Data Cleaning (MDC). Second phase is classification. Two SVM classifiers are used here. First one is the multi-SVM used for predicting the thyroid disease type ie, Euthyroid, Hypothyroid, Sub-clinical hypothyroid and Sub-clinical Hyperthyroid.

Second classifier used is a binary SVM for predicting the chances of goitre.

Qiao pan et. al in [4] have proposed a classification method based on random Forest. Principal Component Analysis is used for Feature dimension reduction. Classification model is set up using C4.5 decision tree. The reduced feature set is classified using random forest ensemble algorithm. Final results are obtained by K 10-fold cross validation.

V. Prasad et al in [2] have proposed hybrid architecture for identifying the thyroid disease and the disease type. Rough data sets theory (RDS) is used here for finding the missing values in the input. Here, a String Matching System is proposed with Particle Swarm optimization and Artificial Bee Colony Optimization. The system is further enhanced with Rule Based System.

Most of the methods classify the disease into three classes: hyper, hypo and normal. The main difficulty with Support Vector Machine is selection of parameters. This paper proposes Particle Swarm Optimization to optimize the SVM parameters.

III. PROPOSED SYSTEM

Proposed method has two phases. First phase is the training phase. Thyroid gland dataset is taken from UCI Machine Learning Repository [14]. Hyper thyroid and Hypo thyroid data are taken. Dataset consists of 21 attributes. Most of the attributes are numeric or Boolean valued. The first step in training phase is pre-processing. The records in the dataset with missing values are eliminated. Support Vector Machine (SVM) is using for classification. For improving the classification accuracy Particle Swarm Optimization (PSO) can be used to optimize the SVM parameters. A user interface is giving to the user in the testing phase. User can enter values for the given attributes according to their test result. KNN imputation is using to approximate missing values in the user inputs. Finally, the disease type is diagnosed and provides disease description along with healthy advices. Classification can be done into four categories as given below:

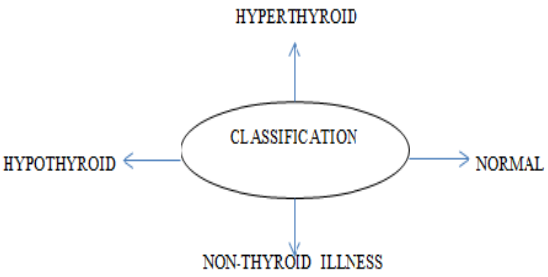

Fig.1. Classification

Architecture of the proposed system is given below:

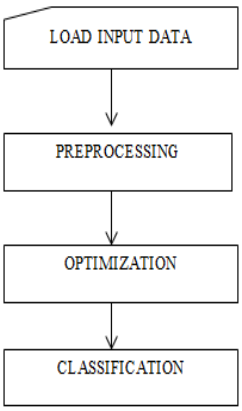

Fig.2. Proposed system training phase

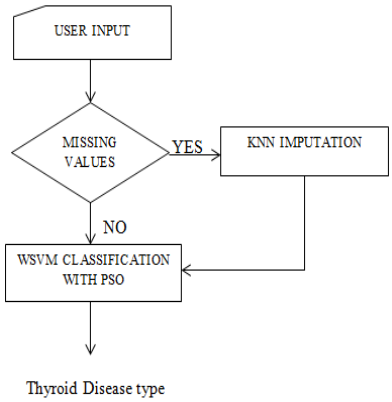

Fig.3. Proposed system testing phase

A. *KNN Imputation*

KNN is an algorithm that is useful for matching a point with its closest K-neighbors in a multi-dimensional space. The steps for imputation are:

1. Add the user inputs with missing value in to the dataset.
2. Sets the value for K.
3. KNN considers K nearest neighbors from the dataset.
4. Value of the missing field is approximated by the corresponding field values of the neighbors.

B. *Support Vector Machine*

SVM is a supervised algorithm and can be used for both classification and regression. SVM plots the data items in the dataset in space as a point. If the dataset consists of n features then the features are plotted in n-dimensional space. Classification is performed by finding a hyper plane that differentiates two classes. SVM is binary classifier but it can be used for the training and classification of multiple classes. Support vectors are the co-ordinates of individual observations. SVM ignores outliers and find the hyperplane that has maximum marginal distance from the two classes.

There are different approaches for multiclass SVM classification. Two main approaches are one-versus-one and one-versus-rest approach. One-versus-all approach is used here. Four binary SVM classifiers are used here for classification. Hyperplane is determined by the kernel function. For non-linear classification, usually used kernel functions are polynomial, radial basis function (RBF), and sigmoid function. RBF is used here as the kernel function. The value of RBF depends only on the distance from the origin, ie.

$$\Phi(X) = \| \Phi(X) \| \quad (1)$$

Parameter selection is very important in SVM classification. Main parameters are Error penalty parameter (C) and Gamma. If C is small then the decision surface will be smooth. When it is high training samples can be classified correctly. The model behavior is sensitive to gamma.

C. Particle Swarm Optimization

PSO is inspired by social behavior among individuals like the birds flocking or fish grouping and was proposed by James Kennedy and Russell Eberhart in 1995. It consists of a swarm of particles that search for the best position according to best solution. During iterations, each particle moves in the direction of its best personal position and global position. PSO combines self-experience with social-experience. Basic concept of PSO is to accelerate each particle toward its personal best and global best locations. Each particle is associated with coordinates in the solution space. These coordinates represent the best solution that has so far achieved by that particle. This is the personal best. Another value considered by PSO is global best which is the best value in the neighborhood of that particle. The basic principles [13] of swarm intelligence are :

1. *Proximity principle*

The population should have the ability to perform simple space and time computation.

2. *Quality Principle*

The population should respond to the quality factors in the environment.

3. *Principle of Diverse Response*

The population should not perform its activities through narrow channels.

4. *Principle of Stability*

The behavior of the population should not change with environmental changes.

5. *Principle of Adaptability*

The population should be adaptable with the computational price.

D. PSO-SVM Classification

For non-linear SVM classification two parameters are considered. The values of these parameters have greater influence in the classification accuracy of SVM. The parameters are Error penalty parameter(C) and Gamma. Error penalty parameter controls the complexity of model and approximate error. Gamma restricts the complexity degree of the optimal separating plane. In order to improve the classification accuracy these parameters can be optimized using PSO. Proposed algorithm is as follows:

1. Initialization
   Set the values for PSO parameters for the training set. The training set forms the swarm.
2. Establish PSO population and initialize speed and location of each particle.
3. Train SVM according to its particle values.
4. Calculate the corresponding fitness function using the equation:
   Correctly classified samples/ total samples in the swarm or training set
5. Do circulation until solution doesn't change anymore or reach maximum number of iterations.
6. Update the location and speed of each particle.
7. Selects parameter values from the particle's best solution.

Equation for velocity update is given below:

$$V_i(k+1) = \omega V_i(k) + C_1 R_1 (PBest_i - X_i(k)) + C_2 R_2 (GBest_i - X_i(k)) \quad (2)$$

Where, ω, C_1, C_2 are constants. R_1, R_2 are random variables. Equation for position update is as follows:

$$X_i(k+1) = X_i(k) + V_i(k+1) \quad (3)$$

Training set forms the swarm here. Swarm size is equal to the size of the dataset. C_1, C_2 are taken as 2. Here, PSO is used for the optimization of two parameters, so that problem dimension will be 2. Maximum number of iteration is set to 100. After completing the specified number of iterations, parameters are selected from the best position of the particle in the swarm.

IV. Conclusion

By using the proposed method thyroid disease can be classified and the disease type can be diagnosed. The system can also give disease description along with healthy advices that may help the users. KNN imputation is used to approximate the missing values of user input and this process can improve classification accuracy. The proposed method can be applied in health care industry. By diagnosing the disease

type along with healthy advices can eliminate thyroid replacement therapy and thyroid removal up to an extent.

ACKNOWLEDGMENT

I take this opportunity to express my sincere gratitude to all respected personalities who had guided, inspired and helped me in the successful completion of this paper. First and foremost, I express my thanks to The Lord Almighty for guiding me in this endeavor and making it a success.

REFERENCES

[1] Jamil Ahmed Chandio, M. Abdul Rehman Soomrani, "TDTD: Thyroid disease type diagnostics", Intelligent Systems Engineering, 2016 International Conference .

[2] V. Prasad, T. Sreenivasa Rao, M. Surendra Prasad Babu "Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms", Springer, 2015.

[3] K.Geetha , Capt. S. Santhosh Baboo, "An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method", Global Journal of Computer Science and Technology: E Network, Web & Security Volume 16 Issue 1 Version 1.0 Year 2016.

[4] Qiao Pan, Yuanyuan Zhang, Min Zuo, Lan Xiang, Dehua Chen,"Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest", Information Technology in Medicine and Education (ITME), 2016 8th International Conference.

[5] Khushboo Chandel, Veenita Kunwar , Sai Sabitha , Tanupriya Choudhury "A comparitive study on thyroid disease detection using k-nearest neighbor and naïve bayes classification techniques" Springer, 2017.

[6] Feyzullah Temurtas, "Acomparitive Study on Thyroid Disease Diagnosis using Neural Networks", Elsevier,2009.

[7] Fatemeh Saiti, Afsaneh Alavi Naini, Mahdi Aliyari Shoorehdeli, Mohammad Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM", International Conference on Bioinformatics and Biomedical Engineering, 2009.

[8] Jianning Chi, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, Mark Eramian, "Thyroid Nodule Classification In Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network", Springer, 2017.

[9] Qin Yu, Tao Jiang, Aiyun Zhou, Lili Zhang, Cheng Zhang, Pan Xu, "Computer Aided Diagnosis of malignant or benign thyroid nodes based on ultrasound images", Springer, 2017.

[10] Hiroshi de Silva, A. Shehan Perera, "Missing Data Imputation Using Evolutionary K-Nearest Neighbor Algorithm for gene expression data", International Conference on Advances in ICT for emerging Regions, 2016.

[11] Tahani Aljuaid, Sreela Sasi, "Proper imputation techniques for missing values in data sets", International Conference on Data Science and Engineering, 2016.

[12] M. A. Hearst, S T Dumais, E.Osuna, J. Paltt, B. Scholkpof, "Support Vector Machines", IEEE Intelligent Systems and their Applications, 1998.

[13] J. Kennedy, R. Eberhart, " Particle Swarm Optimization", IEEE International Conference on Neural Networks, 1995.

[14] http://archive.ics.uci.edu/ml/datasets/thyroid+disease